A. Training and Inference Algorithms

In this section, we present detailed training and inference algorithms of the proposed DiffSal framework.

Training. In the training phase, we perform the diffusion process that corrupts ground-truth saliency maps S_0 to noisy maps S_t, and train the Saliency-UNet to reverse this process. Algorithm 1 provides the overall training procedure.

Inference. Algorithm 2 summarizes the detailed inference process of the proposed DiffSal. The parameter *steps* denotes the number of iterative denoising steps. Specifically, at each sampling step, the Saliency-UNet takes as input random noisy maps or the predicted saliency maps of the last sampling step and outputs the estimated saliency maps of the current step. We then adopt DDIM to update the heatmaps for the next step.

Algorithm 1: DiffSal Training

Input: frames: I, audio: A, T, gt maps: S_0

repeat

  $\mathbf{f}_v = \textbf{VideoEncoder}(I)$;

  $\mathbf{f}_a = \textbf{AudioEncoder}(A)$;

  $t \sim \text{Uniform}(1, ..., T)$;

  $S_t = \sqrt{\bar{\alpha}_t} S_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \epsilon \in \mathcal{N}(0, \boldsymbol{I})$;

  Take gradient descent step on $\Delta_\theta \| g_\psi(S_t, t, \mathbf{f}_a, \mathbf{f}_v) - S_0 \|_2^2$

until *converged*

Algorithm 2: DiffSal Inference

Input: frames: I, audio: A, $steps$, T

Output: predicted saliency map: S_{pred}

$\mathbf{f}_v = \textbf{VideoEncoder}(I)$;

$\mathbf{f}_a = \textbf{AudioEncoder}(A)$;

$S_t \sim \mathcal{N}(0, \boldsymbol{I})$;

$\text{times} = \text{Reversed}(\text{Linespace}(-1, T, steps))$;

$\text{time}_{\text{pairs}} = \text{List}(\text{Zip}(\text{times}[:-1], \text{times}[1:]))$;

for t_{now}, t_{next} **to** $\text{time}_{\text{pairs}}$ **do**

  $S_{pred} = g_\psi(S_t, t_{now}, \mathbf{f}_a, \mathbf{f}_v)$

  $S_t = \text{DDIM}(S_t, S_{pred}, t_{now}, t_{next})$

B. Supplementary Experiments

This section continues the analysis of DiffSal's components, evaluates DiffSal's performance on three video datasets, and presents visualization results.

B.1. Further analysis of DiffSal

Analyzing the Performance of DiffSal using Different Video Encoders. We conduct experiments within DiffSal using video encoders employed in other SOTA works, *e.g.*, the 3D ResNet in STAViS and the S3D in CASP-Net, as illustrated in the table below. In comparison to Table 7, DiffSal (w/ S3D) surpasses CASP-Net, while DiffSal (w/ 3D ResNet) also outperforms STAViS. This highlights the superiority of our diffusion model-based framework under the same encoders and affirms DiffSal's adaptability to various types of encoders.

Method	AVAD		ETMD		Coutrot1	
	CC ↑	SIM ↑	CC ↑	SIM ↑	CC ↑	SIM ↑
DiffSal(w/ 3D ResNet)	0.632	0.471	0.583	0.441	0.521	0.417
DiffSal(w/ S3D)	0.708	0.541	0.637	0.492	0.578	0.469
DiffSal(w/ MViT)	**0.738**	**0.571**	**0.652**	**0.506**	**0.638**	**0.515**

Table 7. Compare the performance of DiffSal using different video encoders.

Analyzing the Number of Multi-modal Attention Modulation Stages. The decoder part of the Saliency-UNet is configured with four stages by default. Figure 6 shows the impact of varying the number of multi-modal attention modulation stages on task performance across the AVAD and ETMD datasets. Notably, the most optimal performance is achieved when the number of multi-modal attention modulation stages is set to **4**. These results imply that Saliency-UNet benefits from progressively fusing audio and video features at multiple scales.

Visualizing Key Audio-Visual Activities. Figure 7 illustrates the key audio-visual activity features learned by the multi-modal interaction module during the generation of the saliency maps in DiffSal. It is obvious that the highlighted key audio-visual activity regions correspond well to the sound sources in the frame. For example, the DiffSal model can focus on the main speaker in two-person dialog scenes with the help of sound, and attend to the position of musical instruments in playing scenes. This further confirms the ability of the proposed multi-modal interaction module to capture key audio-visual activity regions, which in turn enhances saliency prediction performance.

B.2. Comparison with Video Saliency Prediction Methods

For comprehensive validation, the performance of the video-only version of the DiffSal model is analyzed on three commonly used video datasets: DHF1k [53], Hollywood2 [36], and UCF-Sports [41]. (i) DHF1k comprises 600 training videos, 100 validation videos, and 300 testing videos, all with a frame rate of 30 fps. The DiffSal model can only be evaluated on the validation set of DHF1k due to unavailable annotations of the test set, following [28, 35]. (ii) Hollywood2 consists of 1707 videos extracted from 69 movies, with 12 categorized action classes. For training, 823 videos

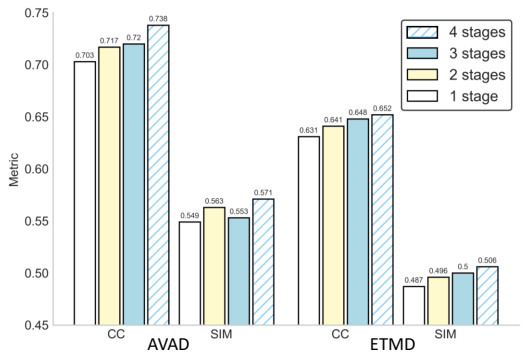

Figure 6. Analyzing the effect of varying the number of multi-modal attention modulation stages on the AVAD and ETMD datasets.

are used, and for testing, 884 videos are utilized. (iii) UCF-Sports contains 150 videos, with 103 for training and 47 for testing. These videos are collected from broadcast TV channels and cover 9 sports, including diving, weightlifting, and horse riding.

Table 8 shows a comparison of the video-only version of DiffSal method against existing state-of-the-arts, including TMFI-Net [64], TinyHD-S [28], and STSANet [55], on three video datasets. Our approach advances the most state-of-the-art methods by an evident margin on DHF1k and Hollywood2 and achieves good performance on UCF-Sports. Compared to TMFI-Net, the CC performance of DiffSal improves from 0.524 to 0.533 on DHF1k and from 0.739 to 0.765 on Hollywood2, respectively. As for the number of parameters and computational complexity of the model, DiffSal has the highest number of parameters, but only about half the computational complexity of the second place TMFI-Net.

The size of the UCF-Sports dataset is minimal compared to the DHF1k and Hollywood datasets, with only 150 videos. Training on the UCF-Sports dataset causes DiffSal with more parameters to be difficult to converge completely, and only achieves a sub-optimal state. While other models with less number of parameters are easier to fully optimize on the UCF-Sports dataset. These experimental results show that the DiffSal model achieves a balance between performance and computational complexity.

B.3. More Qualitative Analysis

Figures 8 and 9 show the performance of DiffSal in diverse real-world scenarios, respectively. These visualizations demonstrate that DiffSal's predictions are much closer to the ground-truth maps, whereas the CASP-Net and STAViS methods struggle to predict the accurate saliency regions.

C. Limitations and Future Work

While DiffSal provides an effective and generalized diffusion-based approach for audio-visual saliency prediction, it also increases the number of parameters and computational complexity of the model. Exploring ways to lighten the model can further enhance its applicability, *e.g.*, to edge devices with limited computational power.

Figure 7. Visualizing the key audio-visual activity features learned by multi-modal interaction module when generating saliency maps. Each pair of pictures shows the frame of the sounding object in the scene (left) and the key audio-visual activity area overlaid (right).

Table 8. Comparison with state-of-the-art methods on three video datasets. **Bold** text in the table indicates the best result, and underlined text indicates the second best result. Our DiffSal is comparable to the previous state-of-the-arts.

Method	#Params	#FLOPs	DHF1k				Hollywood2				UCF-Sports			
			CC ↑	NSS ↑	AUC-J ↑	SIM ↑	CC ↑	NSS ↑	AUC-J↑	SIM ↑	CC↑	NSS↑	AUC-J↑	SIM↑
TASED-Net$_{ICCV'2019}$[37]	21.26M	91.80G	0.440	2.541	0.898	0.351	0.646	3.302	0.918	0.507	0.582	2.920	0.899	0.469
UNIVSAL$_{ECCV'2020}$[18]	3.66M	14.82G	0.431	2.435	0.900	0.344	0.673	3.901	0.934	0.542	0.644	3.381	0.918	0.523
ViNet$_{IROS'2020}$[30]	31.10M	115.26G	0.460	2.557	0.900	0.352	0.693	3.730	0.930	0.550	0.673	3.620	0.924	0.522
VSFT$_{TCSVT'2021}$[35]	14.11M	60.16G	0.462	2.583	0.901	0.360	0.703	3.916	0.936	0.577	-	-	-	-
ECANet$_{NeuroComputing'2022}$[59]	-	-	-	-	-	-	0.673	3.380	0.929	0.526	0.636	3.189	0.917	0.498
STSANet$_{TMM'2022}$[55]	-	-	-	-	-	-	0.721	3.927	0.938	0.579	0.705	**3.908**	0.936	0.560
TinyHD-S$_{WACV'2023}$[28]	3.92M	40.22G	0.492	2.873	0.907	0.388	0.690	3.815	0.935	0.561	0.624	3.280	0.918	0.510
TMFI-Net$_{TCSVT'2023}$[64]	53.41M	305.15G	0.524	3.006	0.918	**0.410**	0.739	**4.095**	0.940	0.607	**0.707**	3.863	**0.936**	**0.565**
Our(DiffSal)	70.54M	161.06G	**0.533**	**3.066**	**0.918**	0.405	**0.765**	3.955	**0.951**	**0.610**	0.685	3.483	0.928	0.543

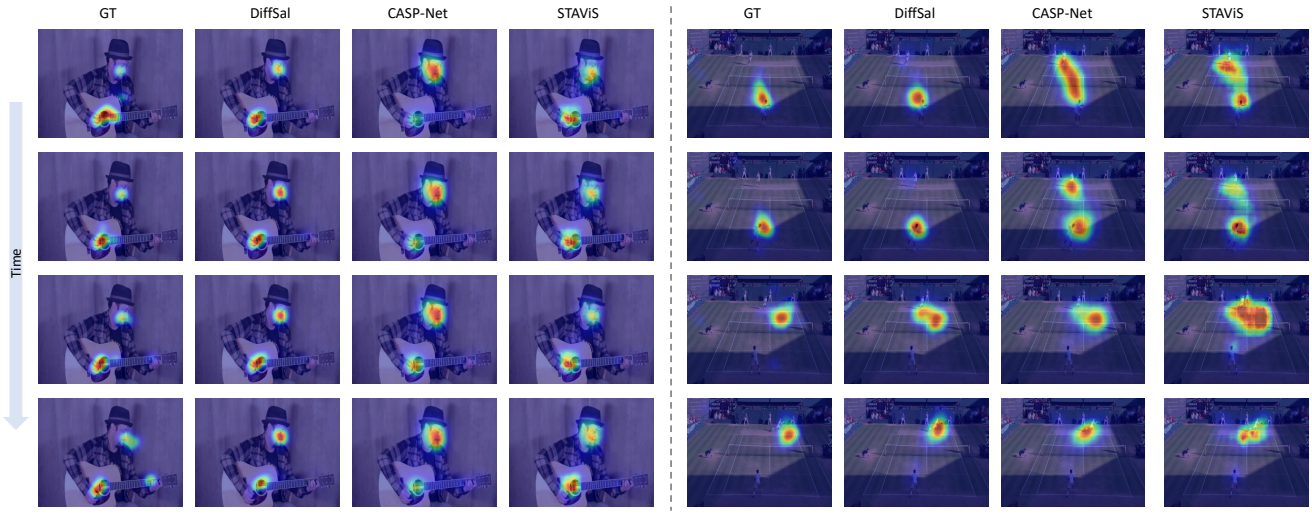

Figure 8. Comparison of visualized saliency maps from the ground-truth, our DiffSal, and previous state-of-the-art CASP-Net and STAViS.

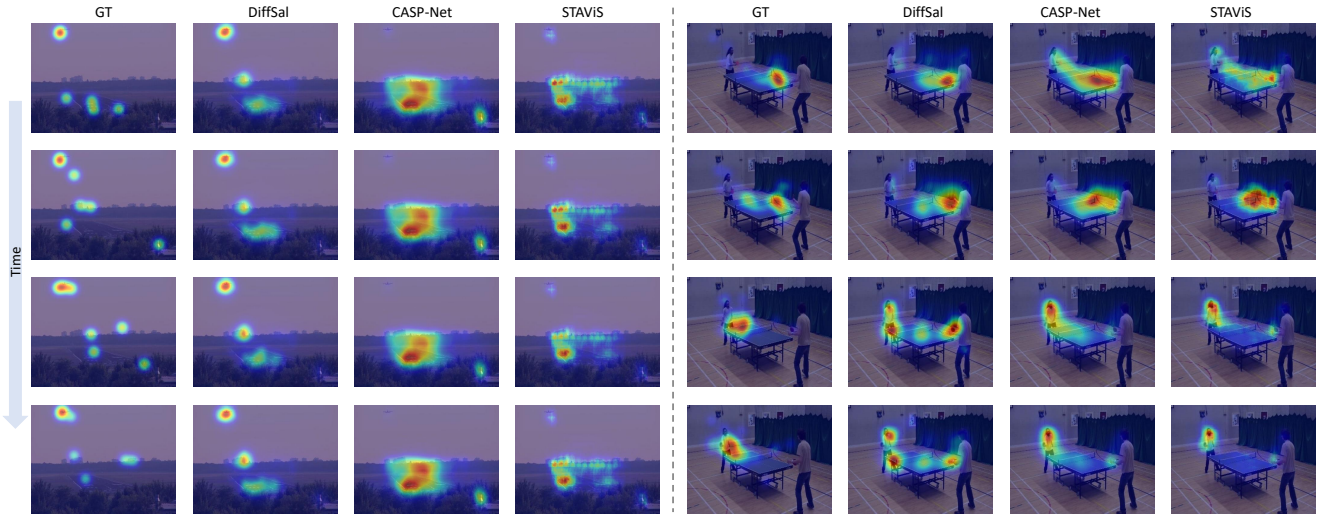

Figure 9. Comparison of visualized saliency maps from the ground-truth, our DiffSal, and previous state-of-the-art CASP-Net and STAViS.